

nuts and bolts of DNA sequencing approaches and bioinformatic tools

Dionysios A. Antonopoulos

Institute for Genomics and Systems Biology
Biosciences Division
Argonne National Laboratory

August 7, 2012

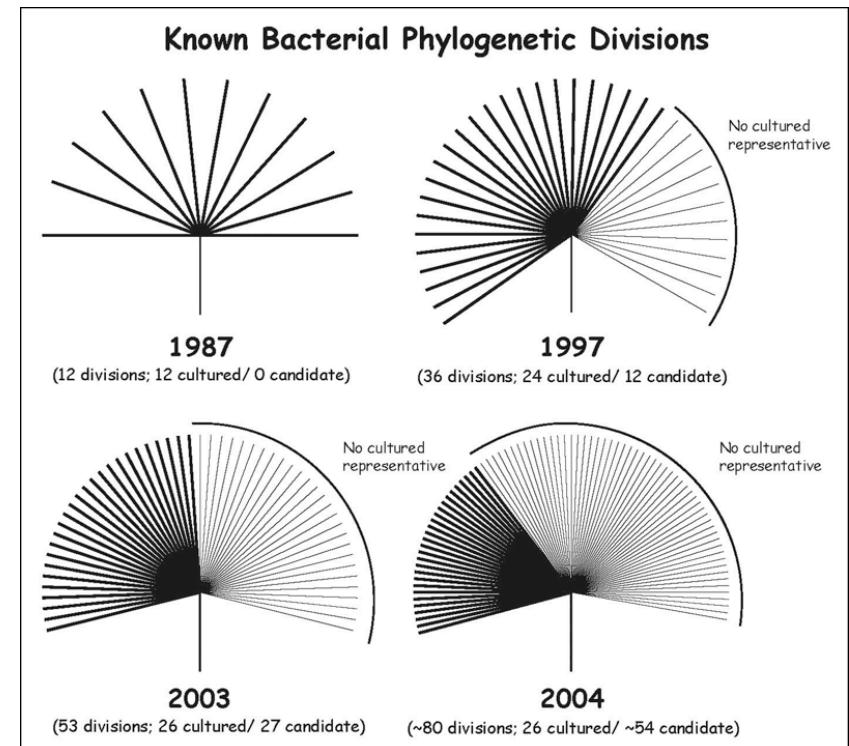
Metagenomics

- Basic concept
- *an approach for obtaining the total gene content of a microbial community in order to understand its function*
- Day-to-day reality of metagenomics
- *has been used interchangeably to describe 16S rRNA-based inventories and whole shotgun approaches*
- *different questions can be addressed (sometimes exclusively) and drastically different hardware needed (both in terms of wetlab and computational)*



How we study microbial communities

- Mode for the past 25 years has been single gene as proxy for whole organism (16S rRNA gene inferred)
- *New insights on extent of uncultivable world*
- *Wide community adoption (i.e. accessible technology and analytical tools)*



Fox (2005) ASM News. 71:6-7



Tools

- Hardware equations for handling 16S rRNA data:

Sanger



454



Illumina



+



-or-



Tools

- Database and software options for handling 16S rRNA data:

DATABASES FOR rRNA SEQUENCES



ANALYSIS PLATFORMS FOR rRNA SEQUENCES



The concept – bookkeeping and accounting

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	# QIIME v1.2.0 OTU table														
2	#OTU ID	FA1	FA10	FA11	FA12	FA2	FA3	FA4	FA5	FA6	FA7	FA8	FA9	FGB1	FGB10
3	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
4	1	0	0	0	3	0	0	0	0	0	2	0	0	0	0
5	2	2	0	0	0	2	0	1	0	1	1	1	1	0	0
6	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	4	0	0	0	0	0	1	0	0	0	0	0	0	0	0
8	5	0	0	0	0	0	0	0	0	0	0	0	0	21	0
9	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	7	0	0	0	0	0	0	0	3	0	0	0	0	0	0
11	8	4	0	0	0	0	0	0	0	0	0	0	0	7	0
12	9	0	0	0	0	0	1	0	0	0	0	0	0	0	0
13	10	0	0	0	0	0	0	0	0	0	0	0	0	1	0
14	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	12	0	0	0	0	6	1	0	0	6	0	0	0	15	0
16	13	1	0	0	0	0	0	0	0	0	0	0	0	0	0
17	14	0	0	0	4	0	0	0	0	0	0	0	0	0	0
18	15	0	0	0	0	1	0	1	1	0	0	0	0	0	0
19	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	17	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	18	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	20	0	0	0	0	3	0	0	0	0	0	0	0	0	0
24	21	0	0	0	0	0	0	0	0	0	0	0	0	0	1
25	22	247	0	0	0	381	76	22	7	13	1	0	0	385	0
26	23	0	0	0	0	0	0	0	0	0	0	0	0	0	0

	A	B	C	D	E	F	G	H	I						
1	# QIIME v1.2.0 OTU table														
2	#OTU ID	FA1	FA10	FA11	FA12	FA2	FA3	FA4	FA5	FA6	FA7	FA8	FA9	FA10	Consensus Lineage
3	0	1	0	0	0	0	0	0	0	0	0	0	0	0	Root;Bacteria;Actinobacteria;Actinobacteria
4	1	0	0	0	3	0	0	0	0	0	2	0	0	0	Root;Bacteria;Proteobacteria;Gammaproteobacteria
5	2	2	0	0	0	2	0	1	0	1	1	1	1	0	Root;Bacteria;Proteobacteria;Betaproteobacteria
6	3	0	0	0	0	0	0	0	0	0	0	0	0	0	Root;Bacteria;Firmicutes;"Clostridia";Clostridiales
7	4	0	0	0	0	0	1	0	0	0	0	0	0	0	Root;Bacteria;Proteobacteria;Betaproteobacteria
8	5	0	0	0	0	0	0	0	0	0	0	0	0	0	Root;Bacteria;Acidobacteria;Acidobacteria
9	6	0	0	0	0	0	0	0	0	0	0	0	0	0	Root;Bacteria;Firmicutes;"Clostridia";Clostridiales
10	7	0	0	0	0	0	0	3	0	0	0	0	0	0	Root;Bacteria;Proteobacteria;Deltaproteobacteria
11	8	4	0	0	0	0	0	0	0	0	0	0	0	0	Root;Bacteria
12	9	0	0	0	0	0	1	0	0	0	0	0	0	0	Root;Bacteria;Actinobacteria;Actinobacteria
13	10	0	0	0	0	0	0	0	0	0	0	0	0	0	Root;Bacteria
14	11	0	0	0	0	0	0	0	0	0	0	0	0	0	Root;Bacteria;Chloroflexi;Anaerolineae
15	12	0	0	0	0	6	1	0	0	6	0	0	0	0	Root;Bacteria
16	13	1	0	0	0	0	0	0	0	0	0	0	0	0	Root;Bacteria;Proteobacteria;Betaproteobacteria
17	14	0	0	0	4	0	0	0	0	0	0	0	0	0	Root;Bacteria;Firmicutes;"Clostridia";Clostridiales
18	15	0	0	0	0	1	0	1	1	0	0	0	0	0	Root;Bacteria;Proteobacteria;Betaproteobacteria
19	16	0	0	0	0	0	0	0	0	0	0	0	0	0	Root;Bacteria;Firmicutes;"Clostridia";Clostridiales
20	17	0	0	0	0	0	0	0	0	0	0	0	0	0	Root;Bacteria;Firmicutes;"Clostridia";Clostridiales
21	18	0	0	0	0	0	0	0	0	0	0	0	0	0	Root;Bacteria;Proteobacteria;Gammaproteobacteria
22	19	0	0	0	0	0	0	0	0	0	0	0	0	0	Root;Bacteria
23	20	0	0	0	0	3	0	0	0	0	0	0	0	0	Root;Bacteria;Proteobacteria;Alphaproteobacteria
24	21	0	0	0	0	0	0	0	0	0	0	0	0	0	Root;Bacteria;Firmicutes;"Clostridia";Clostridiales
25	22	247	0	0	0	381	76	22	7	13	1	0	0	385	Root;Bacteria;Proteobacteria;Gammaproteobacteria
26	23	0	0	0	0	0	0	0	0	0	0	0	0	0	Root;Bacteria;Firmicutes



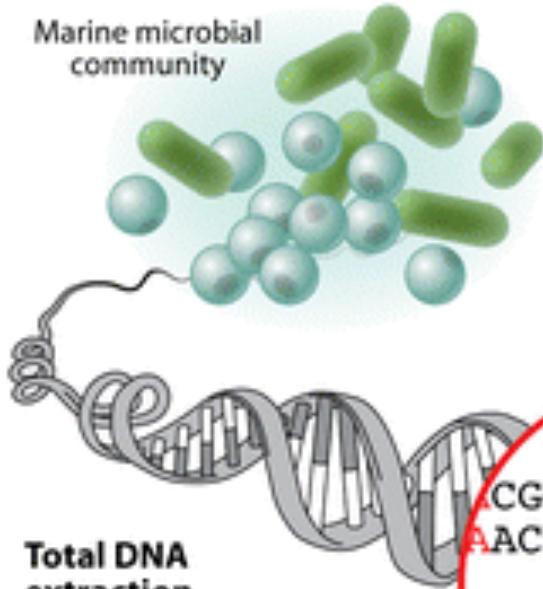
CHALLENGE #1: WORKING IN A DE-CENTRALIZED/ EMERGING FIELD

DATABASES FOR rRNA SEQUENCES



ANALYSIS PLATFORMS FOR
rRNA SEQUENCES





Total DNA extraction

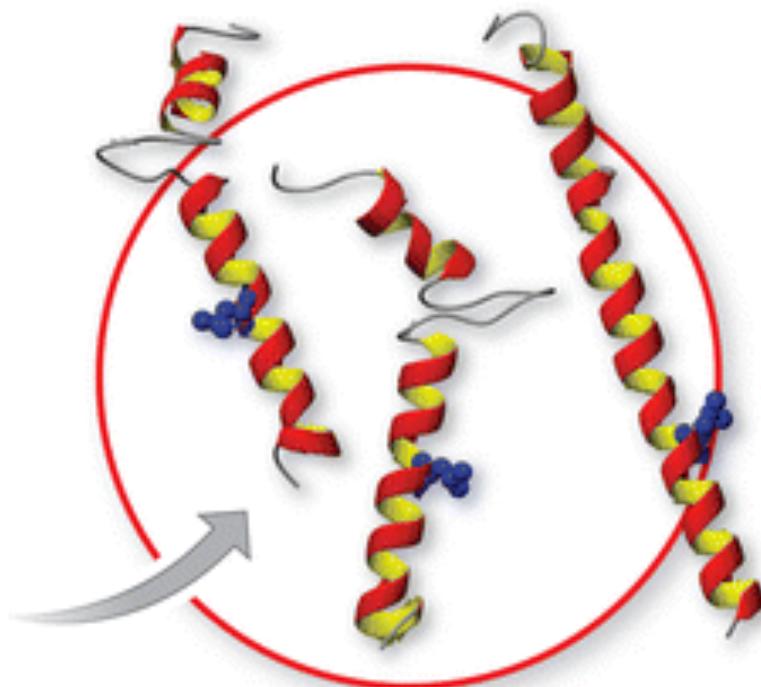
- Environmental single-gene surveys
- Shotgun studies of all environmental genes

ACGCAGAC...
AACTAGCA...

CGAACTAGCATTAA
CGAACGAGCATTAA

DNA sequencing

- Identify common genes within a community
- Identify genome contents favored by current environmental conditions



Protein annotation

Use metagenomics studies as a tool to answer broader ecological or evolutionary questions

A Gilbert JA, Dupont CL. 2011.
R Annu. Rev. Mar. Sci. 3:347–71



The concept – bookkeeping and accounting



	A	B	C	D	E	F
1	function	abundance	avg eValue	avg % ident	avg align len	# hits
2	Alanine racemase (EC 5.1.1.1)	23	-12.88	62.63	60.72	22
3	Alanine racemase, biosynthetic (EC 5.1.1.1)	1	-11	60.34	58	1
4	Alanine racemase, catabolic (EC 5.1.1.1)	2	-12	64.81	54	1
5	Branched-chain amino acid aminotransferase (EC 2.6.1.42)	36	-25.91	65.13	88.47	33
6	Chaperone protein HscA	11	-38.82	66.23	118.45	11
7	Cysteine desulfurase (EC 2.8.1.7)	148	-25.85	63.68	89.29	125
8	Cysteine desulfurase (EC 2.8.1.7), IscS subfamily	9	-41.25	66.83	122	8
9	Cysteine desulfurase (EC 2.8.1.7), NifS subfamily	6	-27.8	60.13	105	4
10	Cysteine desulfurase (EC 2.8.1.7), SufS subfamily	36	-37.39	63.4	117.36	30
11	Cysteine desulfurase CsdA-CsdE (EC 2.8.1.7), main protein	4	-33.67	62.58	120.33	3
12	Ferredoxin, 2Fe-2S	21	-15.11	69.23	53.84	18
13	Iron binding protein IscA for iron-sulfur cluster assembly	4	-24.5	68.35	77	4
14	Iron-sulfur cluster assembly scaffold protein IscU	7	-19.57	73.9	59	7
15	Iron-sulfur cluster regulator IscR	7	-11.67	63.74	55.5	6
16	Probable valine-pyruvate aminotransferase (EC 2.6.1.66)	1	-65	72.12	165	1
17	Valine--pyruvate aminotransferase (EC 2.6.1.66)	1	-78	64.79	213	1
18	2-amino-3-ketobutyrate coenzyme A ligase (EC 2.3.1.29)	32	-47.26	66.79	133.19	26
19	L-threonine 3-dehydrogenase (EC 1.1.1.103)	18	-30.65	68.98	93.35	18
20	Low-specificity L-threonine aldolase (EC 4.1.2.5)	29	-26.24	65.91	88.93	26



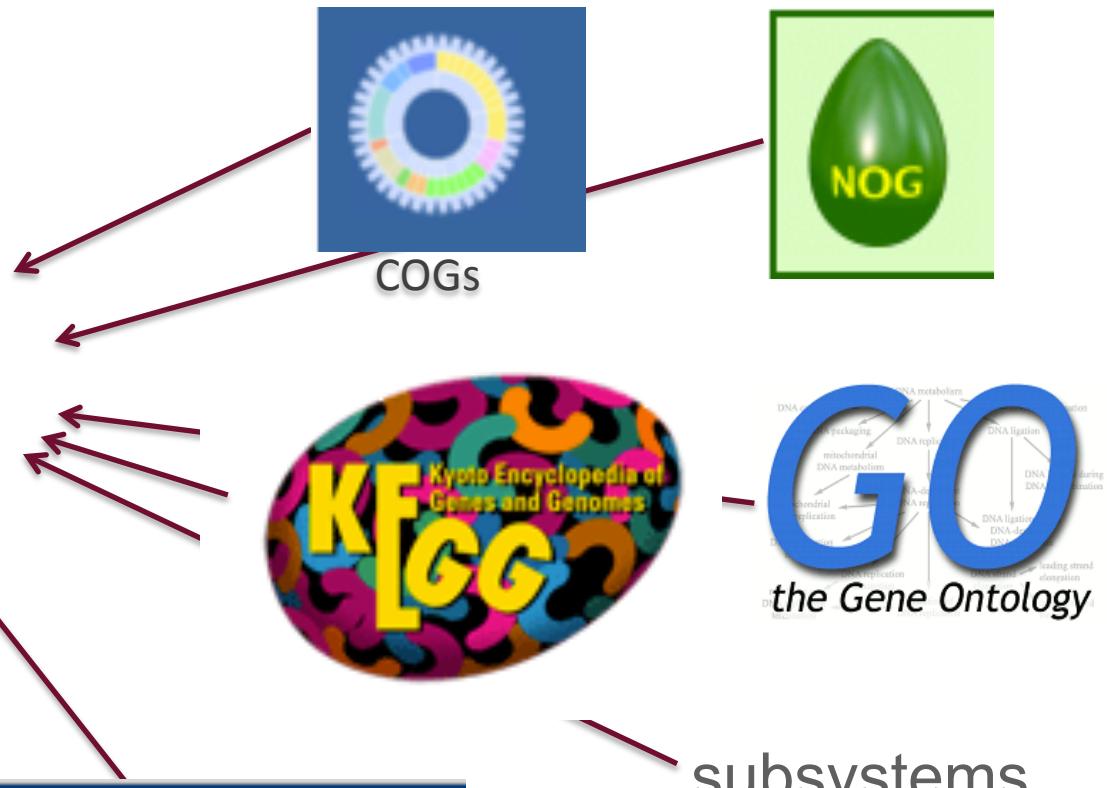


nr

RIBOSOMAL DATABASE PROJECT

silva
comprehensive ribosomal RNA databases

green
genes
16S rRNA gene database and
workbench compatible with ARB
greengenes.lbl.gov



The M5 Initiative



- *Initiative by the Genomic Standards Consortium (GSC); the 5 "M's" in M5 stand for the intersection of:*
 - Metagenomics
 - Metadata
 - MetaAnalysis
 - Models
 - MetaInfrastructure
- In brief, bring together a set of strategic partners investing in various aspects of new and emerging technologies (workflows, grids, clouds, turn-key desktop solutions, etc) to build a next-generation computing landscape.
- *i.e. community-share platform more akin to those built by the physics (e.g. colliders) and astronomy (radiotelescopes) communities.*



It's all about standardization

JUST BECAUSE YOU ARE UNIQUE



DOES NOT MEAN YOU ARE USEFUL

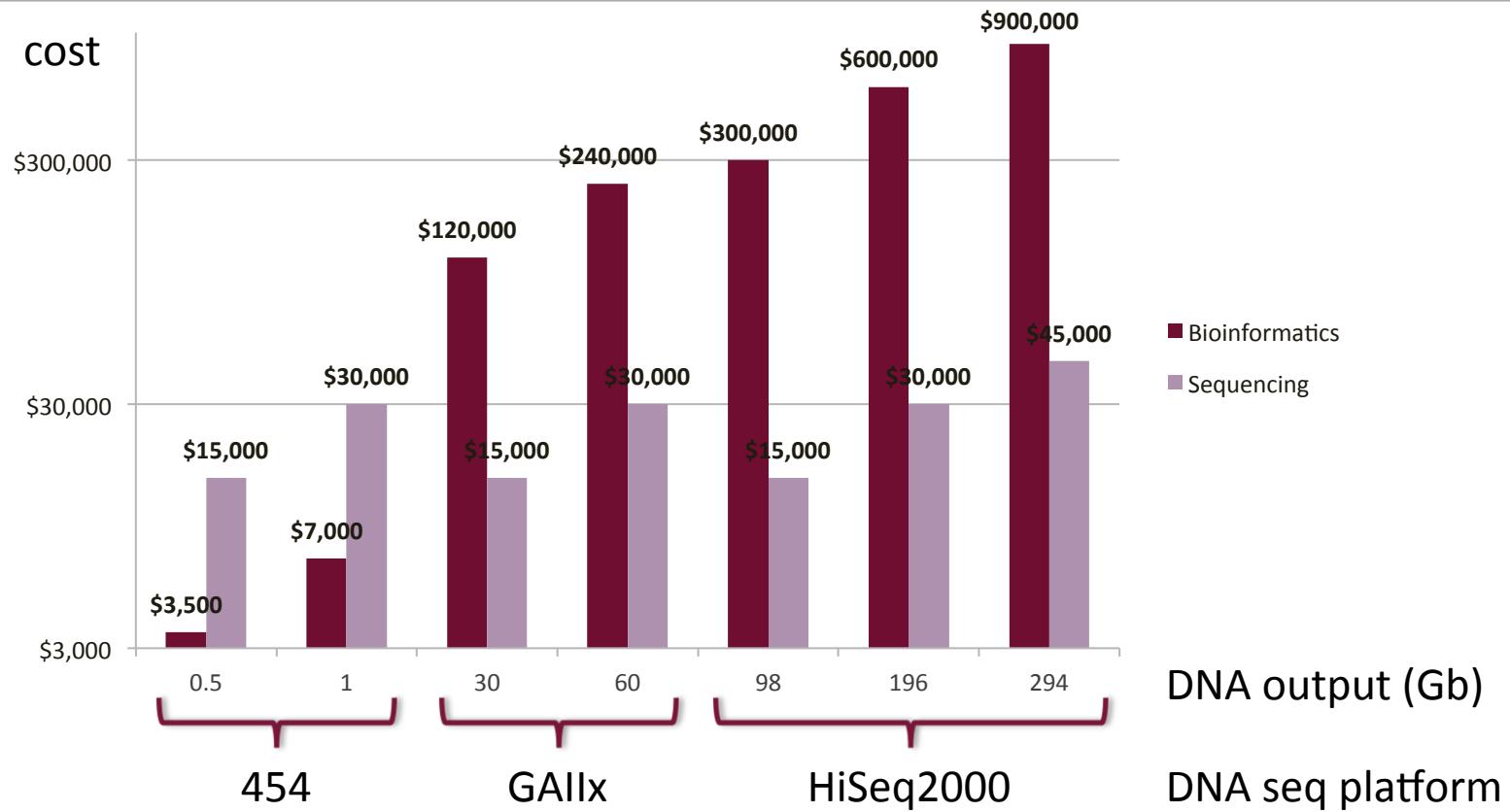
PARENTING.FAILBLOG.ORG



CHALLENGE #2: SCALE OF DATA PRODUCTION HAS CHANGED; IMPACT ON COMPUTATION



Computing cost dominates



Solution to this is to combine:
*novel search algorithms, data reduction strategies, and
cloud computing capabilities*
to process this scale of data

Cost is purely BLASTX on Amazon EC2
Source: Wilkening et al., Proceedings IEEE Cluster09, 2009

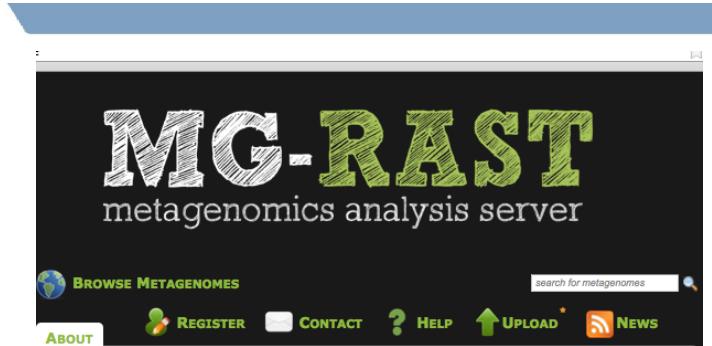
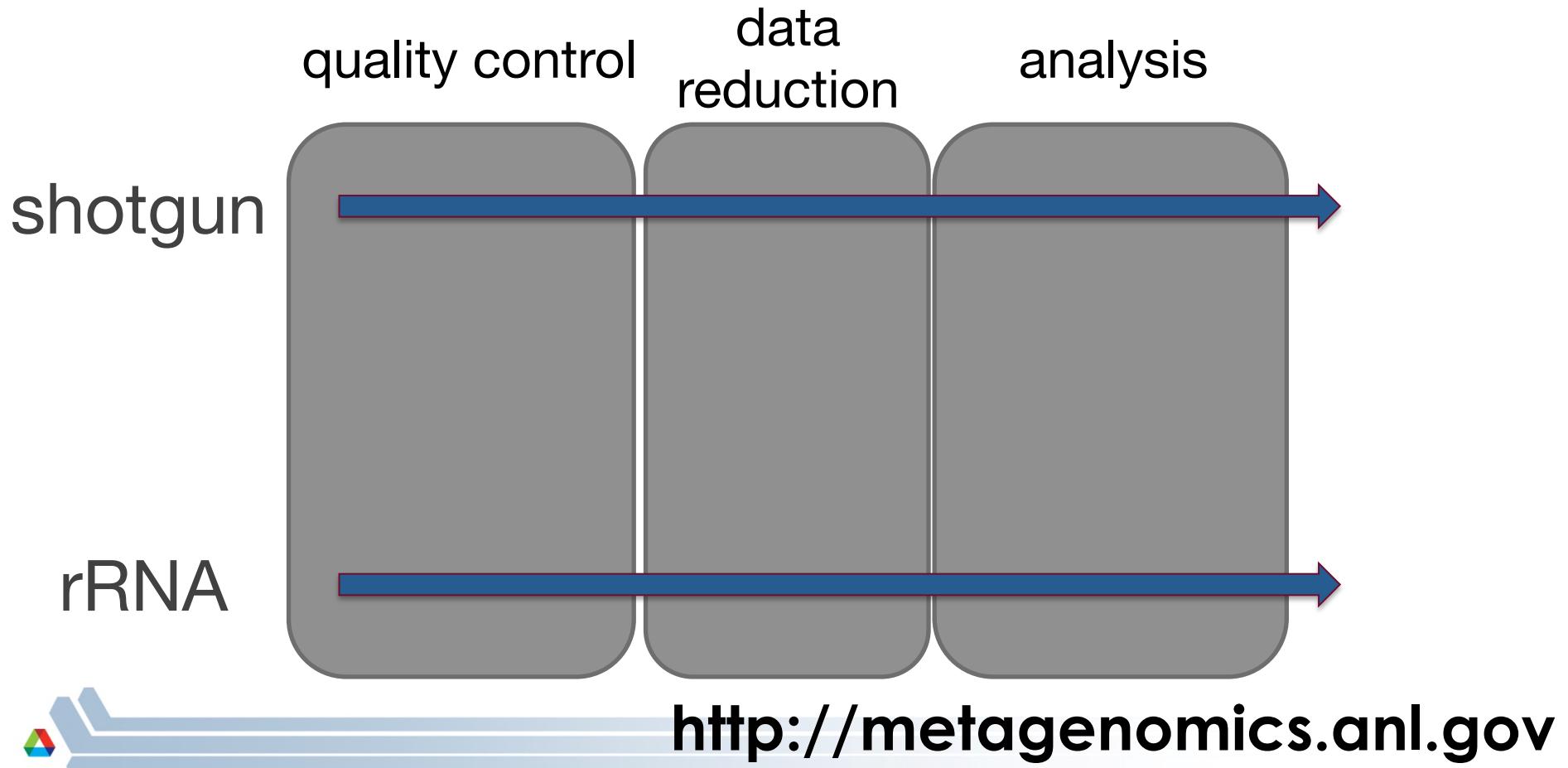


CHALLENGE #3: BIOINFORMATICS IN A TRADITIONALLY DATA (COMPUTATION)-DEFICIENT FIELD



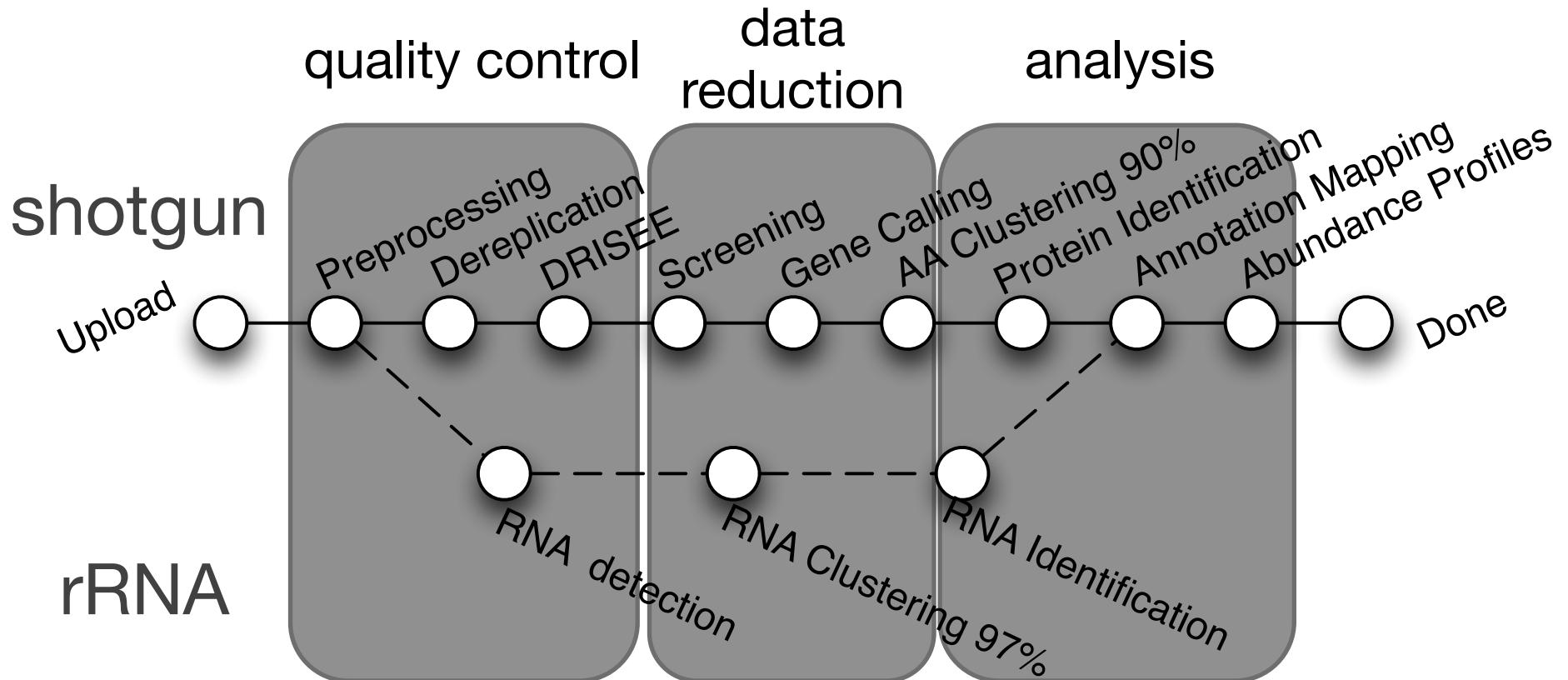
What is MG-RAST?

- Basic concept
- *a web-based resource used to annotate metagenomic datasets and organize the data for queries*



What is MG-RAST?

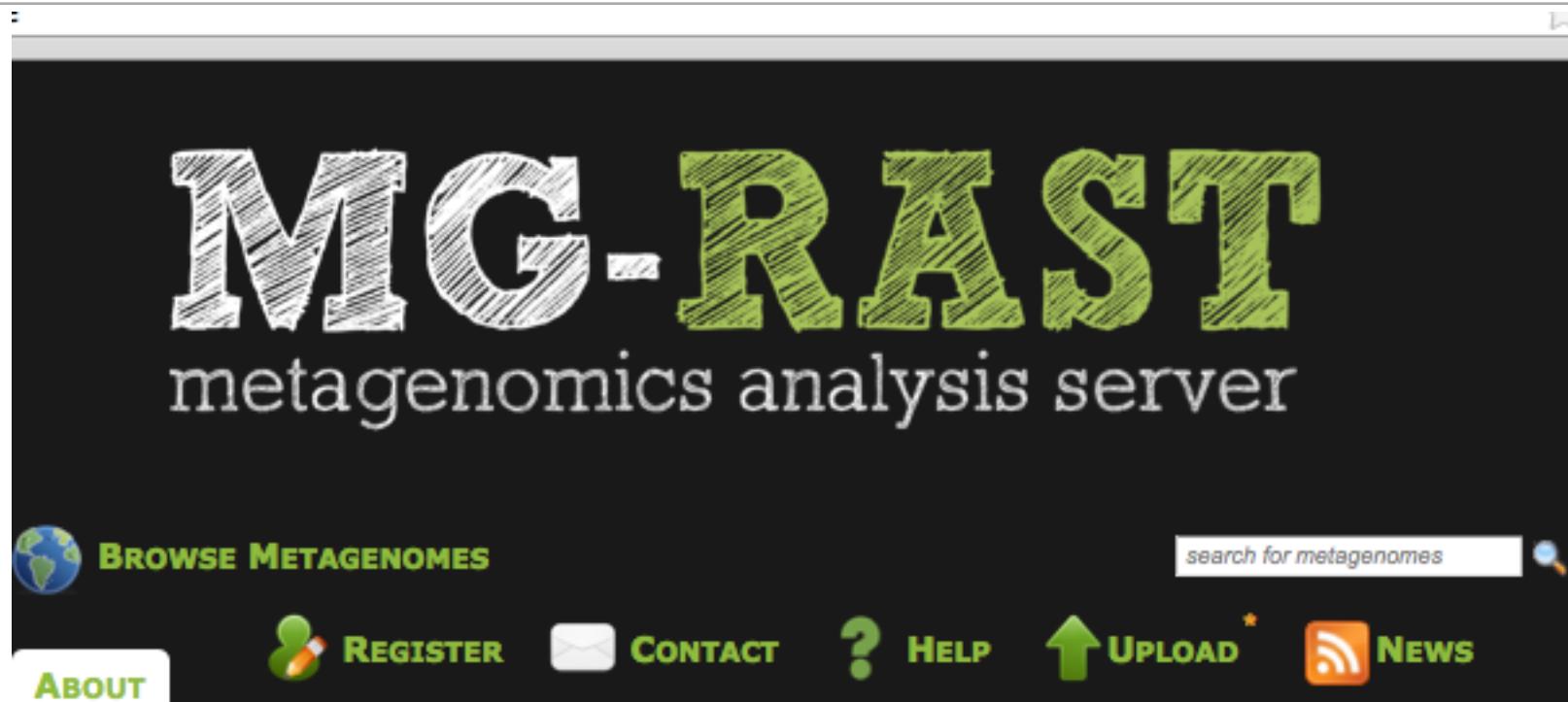
- Basic concept
- *a web-based resource used to annotate metagenomic datasets and organize the data for queries*



<http://metagenomics.anl.gov>



MG-RAST user community contribution



- Rapid growth; 15.11 Tb (as of last night)
- >54,400 datasets from thousands of groups
- *Many labs (not just centers) produce data*



<http://metagenomics.anl.gov>

The concept – bookkeeping and accounting

	A	B	C	D	E	F
1	function	abundance	avg eValue	avg % ident	avg align len	# hits
2	Alanine racemase (EC 5.1.1.1)	23	-12.88	62.63	60.72	22
3	Alanine racemase, biosynthetic (EC 5.1.1.1)	1	-11	60.34	58	1
4	Alanine racemase, catabolic (EC 5.1.1.1)	2	-12	64.81	54	1
5	Branched-chain amino acid aminotransferase (EC 2.6.1.42)	36	-25.91	65.13	88.47	33
6	Chaperone protein HscA	11	-38.82	66.23	118.45	11
7	Cysteine desulfurase (EC 2.8.1.7)	148	-25.85	63.68	89.29	125
8	Cysteine desulfurase (EC 2.8.1.7), IscS subfamily	9	-41.25	66.83	122	8
9	Cysteine desulfurase (EC 2.8.1.7), NifS subfamily	6	-27.8	60.13	105	4
10	Cysteine desulfurase (EC 2.8.1.7), SufS subfamily	36	-37.39	63.4	117.36	30
11	Cysteine desulfurase CsdA-CsdE (EC 2.8.1.7), main protein	4	-33.67	62.58	120.33	3
12	Ferredoxin, 2Fe-2S	21	-15.11	69.23	53.84	18
13	Iron binding protein IscA for iron-sulfur cluster assembly	4	-24.5	68.35	77	4
14	Iron-sulfur cluster assembly scaffold protein IscU	7	-19.57	73.9	59	7
15	Iron-sulfur cluster regulator IscR	7	-11.67	63.74	55.5	6
16	Probable valine-pyruvate aminotransferase (EC 2.6.1.66)	1	-65	72.12	165	1
17	Valine--pyruvate aminotransferase (EC 2.6.1.66)	1	-78	64.79	213	1
18	2-amino-3-ketobutyrate coenzyme A ligase (EC 2.3.1.29)	32	-47.26	66.79	133.19	26
19	L-threonine 3-dehydrogenase (EC 1.1.1.103)	18	-30.65	68.98	93.35	18
20	Low-specificity L-threonine aldolase (EC 4.1.2.5)	29	-26.24	65.91	88.93	26



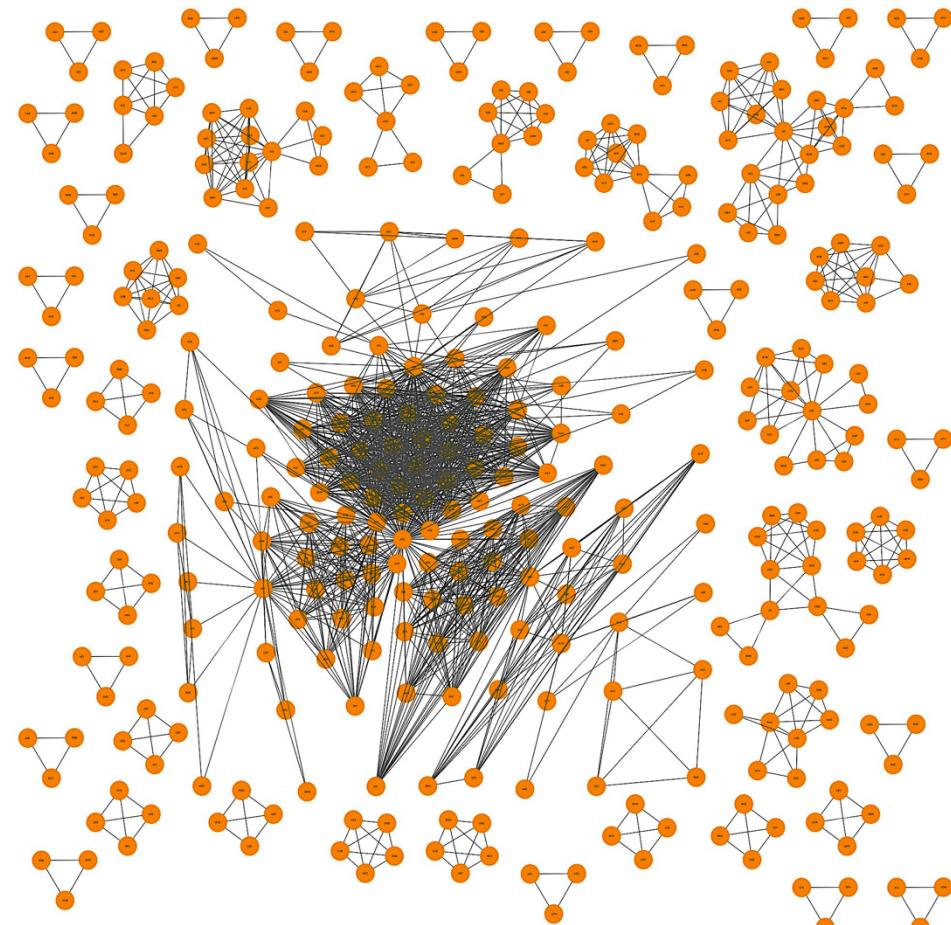
Metadata

- sample_name
- cur_land_use
- cur_vegetation
- cur_vegetation_meth
- previous_land_use
- previous_land_use_meth
- crop_rotation
- agrochem_addition
- tillage
- fire
- flooding
- extreme_event
- horizon
- horizon_meth
- sieving
- water_content_soil
- water_content_soil_meth
- samp_weight_dna_ext
- pool_dna_extracts
- store_cond
- link_climate_info
- annual_season_temp
- annual_season_precpt
- link_class_info
- fao_class
- local_class
- local_class_meth
- soil_type
- soil_type_meth
- slope_gradient
- slope_aspect
- profile_position
- drainage_class
- texture
- texture_meth
- ph
- ph_meth
- tot_org_carb
- tot_org_c_meth
- tot_n
- tot_n_meth
- microbial_biomass
- microbial_biomass_meth
- extreme_salinity
- salinity_meth
- heavy_metals
- heavy_metals_meth
- al_sat
- al_sat_meth
- misc_param



Part of an emerging digital biology community

- *In the current version of MG-RAST, metadata is viewed as the key to providing the intersection between datasets*
- Users (**dots**) sharing pre-publication metagenomes (**edges**)



Source: MG-RAST, 800+ shared metagenomes



Summary

- Sequencing strategies – single gene targeted (proxy for organism) or indiscriminate sequencing (proxy for community function)
- Challenges related to: de-centralized/emerging field, increased scale of data production by next-gen DNA sequencing platforms, and a traditionally data-deficient field
- Solutions in forms of: standardization/exchange of data types, novel data handling strategies, and web-based repositories of organized data
- Metadata, metadata, metadata – standardization and community sharing



Acknowledgements

MG-RAST Team:

- Folker Meyer
- Narayan Desai
- Mark Domanus
- Mark d'Souza
- Elizabeth M. Glass
- Travis Harrison
- Kevin Keegan
- Hunter Matthews
- Sarah Owens
- Tobias Paczian
- Will Trimble
- Andreas Wilke
- Jared Wilkening

Collaborators:

- A. Arkin (Berkeley)
- E. Chang (UChicago)
- Dawn Field (Oxford)
- F.-O. Glöckner (MPI Bremen)
- Jack Gilbert (Argonne)
- Jeff Grethe (CallIT2, CAMERA)
- Sarah Hunter / Guy Cochrane (EBI)
- Rob Knight (Colorado)
- Nikos Kyrpides (DOE JGI)
- Owen White (UMaryland, HMP DACC)

